Against the Grain

Volume 20 | Issue 4 Article 9

September 2008

Not Alone: A Digital Preservation Community

Martha Anderson
Library of Congress, mande@loc.gov

Not Alone: A Digital Preservation Community

by **Martha Anderson** (Director of Program Management, National Digital Information Infrastructure and Preservation Program, Office of Strategic Initiatives, Library of Congress) <mande@loc.gov>

Introduction

In the winter of 2000, revolutionary changes in the digital environment were beginning to take place not only in technology but in the aspects of social interaction and content creation. That time also marked the beginning of a national initiative to develop a strategy for the preservation of a burgeoning body of digital information valuable for scholarship, public policy, and the cultural heritage of the United States. As plans unfolded for the **National Digital Information Infrastructure and Preservation Program (NDIIPP)** led by the **Library of Congress**, the Internet community was engaged in a phenomenon that would be labeled, Web 2.0. While the Web became a platform for collaborative content creation the **National Digital Preservation Program** became a Web of partnerships working together to collect and preserve at-risk digital information.

The **U.S. Congress** recognized the increasingly digital nature of the creative output of the nation. **Public Law 106-554** charged the **Library of Congress** with the leadership of a program to develop a national strategy to preserve digital content for future generations. The legislation mandated a collaborative approach naming a variety of organizations to be included. The first task for the program was to prepare a plan in consultation with numerous representatives from libraries, archives, publishers, producers, technologists, and public and private sector organizations with an interest in digital content. During the consultations, there was wide consensus that no single institution could single-handedly undertake the task of preserving digital content.

Since approval of the plan by **Congress** in 2003, the program has invested over $40 million to engage over 100 partners. Five primary initiatives have undertaken activities that range across basic digital preservation research, the establishment of partnership networks for selection and collection of digital content, the development of software and technical services, collaboration on metadata and format standards, and the study of public policy issues for digital preservation. Full details of the program are presented on the **Digital Preservation** Website, *http://digitalpreservation.gov*.

The goal of the program is to ensure access over time to a rich body of digital content through the establishment of a national network of partners committed to selecting, collecting and preserving at-risk digital information. The approach has been to work iteratively engaging diverse communities to work together on the tools, services, standards and policies

to enable enduring access to valuable digital information. By testing and modeling with a variety of partners, the program has been able to evaluate what a distributed network can accomplish. A valuable outcome of the distributed approach is that program has been able to leverage work across projects to apply lessons learned to similar challenges and provide a wider benefit to the network.

Over the last few years, the **National Digital Preservation Program** has learned that within the larger network of organizations, there are emerging natural networks. These networks are more focused and efficient in working through specific challenges for the preservation of specific content types.

- Thirty-seven national libraries and other international organizations have formed a consortium to work on preserving significant information published on the Internet.
- The motion picture, photography and recorded sound industries are meeting in cross-industry working groups to develop standards for metadata.
- Twenty-five U.S. state libraries and archives are working together to preserve state and local government information.
- A community of public and private sector organizations is working to preserve digital maps and geospatial data.

With the **Library of Congress** serving as a convener and catalyst, the program today has three primary objectives: **identify and collect at-risk born-digital content**, **build and support a national network of preservation partners**, and **develop and adopt technical tools and services for preservation**. Each of these objectives supports each of the others in the efforts to accomplish goals for the long-term.

CONTENT

To date, the preservation partners have identified approximately 130 digital collections as targets for preservation. Most of the content is created originally in digital form but some such as historic maps have been scanned and converted to digital for better access. The collections fall into four general content types: geospatial, text and images, audio and video, and Websites.

Geospatial

Today's maps are born digital and are rich with data critical to land use management, disaster relief, environmental planning and homeland security. Partnerships led by **North Carolina State University**, **North Carolina Center for Geographic Information and Analysis**, **University of California**, Santa Barbara, and **Stanford University** are working to build tools and services, engage partners, and adopt standards for preserving geospatial content. The collection and preservation is focused on information for Congressional cartography studies; state, regional and local government geospatial data (e.g., emergency response assets, jurisdictional boundaries, infrastructure maps), and aerial and satellite imagery, including coastal imagery.

Text and Images

State libraries and archives, a partnership of social science data gathering organizations known as **DataPASS**, several regional consortia in Minnesota, Washington State and Arizona, and photography associations are working on preserving text and images in the form of datasets, public records, journals and reports, graphical presentations, and photographs. These materials represent substantial information investments that have been made by the government, cultural heritage institutions and other segments of society. They include state and local agency records (e.g., court records, vital records, land ownership records), databases containing the results of research and surveys, business records, scholarly journals and digitized cultural heritage content.

"The goal of the program is to ensure access over time to a rich body of digital content ..."

Audio and Video

Television, music recordings and motion pictures are the focus of action by a variety of organizations. There are two television preservation partnerships: a network of public television organizations and a collector of international television broadcasts. The very nature of broadcast distribution makes television and radio one of the most at-risk forms of content. Non-commercial programming from both the U.S. and foreign countries is of particular interest and includes foreign news broadcasts, U.S. television broadcasts, and radio broadcasts. Music producers are working on metadata standards for sound recordings. The **Academy of Motion Picture Arts and Sciences** is leading an industry effort to study archival strategies for digital motion pictures and recommend specifications for image data formats.

Websites

The Web is an increasingly important source of information by and about government, as well as a mirror of the political and social events of our time. Much of the documentation of our daily lives, as well as public discourse and debate, has moved to this new digital landscape in which content appears,

continued on page 36

changes and vanishes at incredible speed. Topics related to critical public policy issues such as public health and medical preparedness, foreign investment and international outsourcing, personal privacy protection and data security, are discussed and explored primarily on the Internet. **California Digital Library**, **University of North Texas**, the **Internet Archive**, the **International Internet Preservation Consortium** and the **Library of Congress** are developing tools and services, standards and collaborative collection approaches for the vast resources published on the Web. Two organizations; **Portico** and **LOCKSS** at **Stanford University** have led the way to preserve electronic journals, distributed via the Web.

Using these collections, the partners in the program have learned about technical requirements for preservation and have developed good practices and tools for various phases of the content life cycle. Technical challenges include the unique characteristics of specific formats within each content type. For the most part, many of the specific challenges arise when issues of description and documentation of the digital objects are considered. Proprietary formats require interaction with commercial software producers and rights-restricted content requires negotiation between rights holders and the stewardship organization. There are natural communities of producers and users within these content types that aid in the discourse and action required for each type of digital information.

A future collaboration will be required across the content types to enable search and discovery within these mixed content type collections. During 2008, the digital preservation partners collaborated on a Web portal linked to highlights of content currently being made available to the partners' communities and/or the public as a demonstration. In the fall of 2008, a meta-discovery prototype will be developed to demonstrate how these preservation collections can be searched. These two exploratory activities will lead to more extensive work on how access requirements may drive preservation decisions for the national collections as a whole. The objective is to develop a better understanding of the effects of access expectations on preservation planning and actions, as well as to understand issues of interoperability for diverse content types.

NETWORK

The program began with the knowledge that the challenge for the preservation of digital content is so great that no single organization can do it alone. The concept of a distributed network of preservation organizations was the foundation of the approach to developing a national strategy. The first investments were for projects to develop partnerships, identify and collect content, and propose technical solutions. The original group of awards went to eight consortia comprising 36 institutions, primarily universities and non-profit organizations. Other early initiatives funded basic research in cooperation with the **National Science Foundation** and an early trial of a proposed preservation architecture.

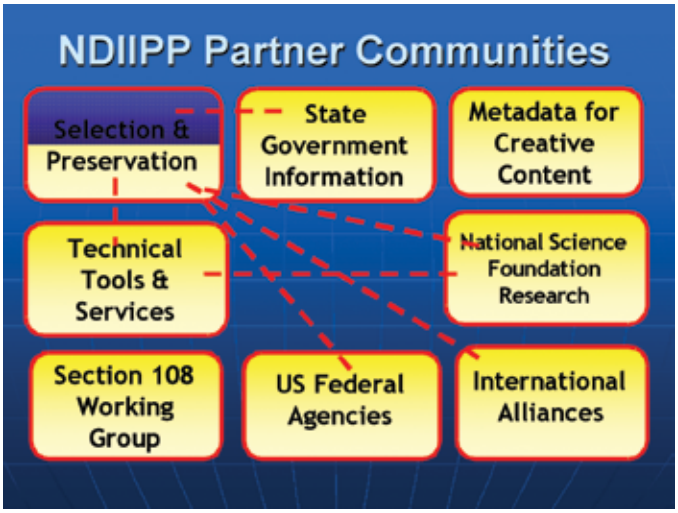
Although each of these initiatives started as independently organized efforts, the **NDIIPP Program Office** at the library was alert for opportunities to leverage work from one initiative for one of the others.

Work done by researchers on the **National Science Foundation/ Library of Congress DigArch** initiative was introduced to projects working on content selection and collection. Often these introductions resulted in expanded work with the researchers learning how to apply their results to specific problems. Another example of leveraging work was the follow up to an early project, the **Archive Ingest and Handling Test**, that tested the ingest and management of identical copies of a Web archive into four institutional systems and the assessment of the effects of that transfer on the integrity of the content. Subsequently the lessons learned in that one-year project were applied by several members of the eight consortia, particularly the lessons learned about the transfer of data and forming partnerships for distributed copies of the data.

The partners meet at least annually to present their work and collaborate on solutions. The meetings facilitate connections between the various projects and the creation of communities of interest around common challenges. The most recent meeting was attended by over 150 participants and included 11 small group sessions to work together on interoperability, Web archiving approaches, incentives for archival deposit, collaborative agreements for building distributed preservation systems and partnerships, strategies for dealing with rights restricted materials, and content transfer and storage options.

Last year, new partners were added with two initiatives, **Preserving State Government Information** and **Preserving Creative America**. The first involves over 25 state libraries and archives working in four projects to develop policies and infrastructures for preserving state and local records, publications, and geospatial data. The second funded eight projects with twenty commercial content producers to target preservation issues across a broad range of creative works, including digital photographs, cartoons, motion pictures, sound recordings and video games. The work is being conducted by a combination of industry trade associations, private sector companies and nonprofits, as well as cultural heritage institutions.

In addition to funding initiatives, the program has been instrumental in addressing policy issues and standards through affiliation with other preservation organizations. The **Digital Preservation Program** supported a two year study group that made recommendations this spring to update Section 108 of the Copyright Law that addresses how libraries may handle materials for the purposes of preservation. A working group of representatives from 11 federal agencies is developing standards for digital still and moving images. The program also maintains international alliances with the **International Internet Preservation Consortium**, the **Digital Preservation Coalition**, and **JISC** in the UK.



Caption: This map of partner communities illustrates the connections between national digital preservation initiatives that are the foundation of a national network of preservation partnerships.

Today the digital preservation partnerships are viewed as a network of networks playing distinct roles as content custodians, preservation service providers, participants in communities of practice and developers of capacity in the forms of funding and education. The plan for the next few years will be to formalize roles and relationships into a national alliance for content stewardship.

TECHNICAL ARCHITECTURE

The technical architecture framework and subsequent tools and services developed through the **Program's** initiatives support and validate the distributed approach.

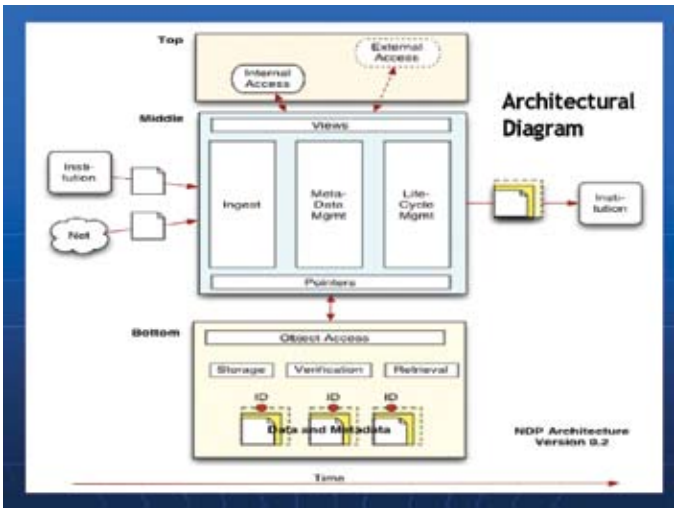
In the first year of planning for the **Digital Preservation Program**, a group of technology experts proposed a three-layer architecture for digital preservation. The lowest layer stores and maintains the data at the bit-level. The middle layer is the stewardship layer that is associated with libraries and archives and provides metadata services. The top layer is the access layer that provides services to render, view and use the content.

These three layers provide a framework for distributed roles to sup-

continued on page 38

Not Alone ...
from page 36

port preservation across time and technological change. Diverse organizations and systems can be configured to leverage the best capabilities in each layer. This view accommodates the highly innovative social networking culture of access on the Internet that encourages creative use of information resources in the top layer and the strategic deliberate action of stewardship organizations that are concerned with the longevity of digital data in the middle layer. Commercial and non-profit data centers provide expertise and services for storage and management at the bottom layer.



Caption: The architectural framework tested and modeled by digital preservation partners in a variety of projects.

One of the earliest program activities was a test of this architectural concept. It set out to simulate the changes over time that digital content would encounter. The change of technology including systems and formats is usually the first concern. However a primary change that should be anticipated in the long-term is change of organizational roles and responsibilities for stewardship. An entity may cease to be capable of maintaining digital content due to loss of resources or redirection of its mission and goals.

The **Archive Ingest and Handling Test** proposed to engage several testing organizations with varied preservation systems and regimes to work with a common archive of diverse digital objects. This archive was transferred to each one and ingested into the local repository or managed storage environment, then exported to another partner for ingest in their system. At every change of environment, the data was examined and evaluated for its integrity and changes brought about by the transfer and ingest actions. A third element but least tested was a migration of formats.

The sample digital archive was donated by **George Mason University**, and the **Library of Congress** conducted the test with **Johns Hopkins**, **Harvard**, **Stanford**, and **Old Dominion** universities. Three partners were academic libraries accustomed to the role of cultural heritage content preservation. The fourth was a computer science department team who provided a good examination of the kinds of tools that proved useful for evaluating and testing the results of change on the digital objects. The archive contained approximately 57,000 files totaling about 12 gigabytes. Although relatively small, it was complex in its mix of formats and metadata.

The archive test proved that different approaches to the same problem can coexist and work successfully and coincidently. We learned which aspects of digital preservation are institution-specific and which aspects are more general. In fact, the library believes that taking several approaches to the same problem is preferable to homogeneity, which risks data corruption or irretrievable loss should the single-system solution fail.

The test also revealed that a data-centric approach to the transfer of content is preferable to a tool-based strategy. Thus, this approach assumes that data will pass among institutions in its original context, to be interpreted by the ingest system of the receiving-preserving institution. Of course, heterogeneous approaches to the same problem can only be successfully guaranteed when networking and cooperation among various institutions exist to the degree necessary to ensure interoperability.

The characteristics of the content play a key role for its longevity. Stewardship organizations must be prepared to fully analyze, characterize and understand the formats and attributes of digital objects in their care. The diversity of digital objects and formats created to date serves to provide further evidence that the stewardship of digital content must be shared across many organizations because one cannot undertake the task alone.

Follow-on technical work across the **National Digital Preservation** partnerships has affirmed the value for open development of tools. As individual partners declared their requirement and intention to develop a specific tool for their local system, others in the partnership offered broader thinking about the requirements and provided able and knowledgeable testers of the tools. Tools for content capture, validation, description, transfer, management, retrieval, storage and access have been developed and adopted by the partners.

A tool that has had wide adoption is the **JHOVE** tool, first developed by **Harvard** and **JSTOR**. The tool's functions analyze and validate digital objects being deposited into a repository. It has proven to be of wider utility to the larger digital preservation community. Today, the **California Digital Library**, **Stanford University Library** and **Portico** are continuing to improve and expand the functionality of **JHOVE** in consultation with the **NDIIPP** digital preservation community and other interested organizations.

Early in the program, partners coping with potentially large volumes of content, identified storage as a pressing need. Not only the provision of storage but an understanding of the principle of redundant copies was required for digital preservation. More than the result of system backup, copies needed to be distributed geographically, organizationally and across diverse systems to be more secure over the long-term. Some of the most productive work has been done with partnerships working with the **San Diego Super Computer Center** and the **Stanford LOCKSS** program to fully test technical solutions for storage and management. The program partners have also held an annual meeting with commercial storage providers to engage in learning from each other about developments in hardware and software that will support the long-term preservation of digital content.

"To date, the preservation partners have identified approximately 130 digital collections as targets for preservation."

Conclusion

Libraries and archives could not predict the revolution in the digital information environment that would erupt with social networking, social tagging, and collaborative content creation. **Flickr**, **Wikipedia**, **Google Maps** and other Web tools and services that changed the nature of content creation and consumption were launched since the **National Digital Information Infrastructure and Preservation Program** was initiated. Congressional legislation authorized the **Library of Congress** to work with other institutions to form a national network of partners dedicated to collecting and preserving important born-digital information. Guided by a strategy of collaboration and iteration, the **Library of Congress** and its partners have been engaged in learning through action. The result is an understanding of appropriate roles and functions for a distributed national network of diverse organizations. Preserving our cultural heritage is not a mission that can be accomplished by a single institution.